Higher Model Robustness by Meta-Optimization for Monocular Depth Estimation

Cho-Ying Wu, Yiqi Zhong, Junying Wang, Ulrich Neumann
University of Southern California

Abstract

Monocular depth estimation is a fundamental task for computer vision. However, its robustness to unseen data, especially images from different datasets that become as adversarial examples, is less explored. This work focuses on indoor monocular depth prediction. We leverage gradient-based meta-learning for higher robustness on zero-shot cross-dataset inference. Unlike the most-studied image classification in meta-learning, depth is pixel-level continuous range values, and mappings from each image to depth vary widely across environments. Thus no explicit task boundaries exist. We instead propose fine-grained task that treats each RGB-D pair as a task in our meta-optimization. We first show meta-learning on limited data induces much better prior (max +29.4%). Using meta-learned weights as initialization, without extra data or information, it shows higher robustness that consistently outperforms baselines.

1. Introduction and Backgrounds

Extending model robustness to open domains is necessary for a practical depth estimator applied to real-world applications. Unlike most previous research for monocular depth estimation [1, 2, 8, 10–16, 20, 30–32] that focuses only on training and testing on a single dataset, we purpose to gear those existing model architecture, either general or dedicated for depth estimation porpose, with higher model generalizability to attain higher robustness.

Meta-Learning principles [7, 25] illustrate an oracle for learning how to learn. Inspired by meta-learning's advantages of domain generalizability, training robust models to achieve better results on unknown domains, usually learned from limited-source data [3, 4, 9, 18], we pioneer to dig into how meta-learning applies to single-image depth estimation. The common meat-learning problem setup follows few-shot multitask settings, where a *task* represents a distribution to sample data from, and most tasks are designed for image classification [9]. Unlike those works, we study a more complex problem of depth estimation: the difficulties lie in per-pixel and continuous range values as outputs, in contrast to global and discrete outputs for image classification. Even for the same environments, images and depth captures can vary greatly, such as adjacent frames for a close-view object can be large room spaces. This observation indicates that our tasks are without clear task boundaries under meta-learning's context [6]. Thus, we propose to treat each training sample as a **fine-grained task**.

We follow the gradient-based meta-learning, which adopts a meta-optimizer and a base-optimizer [3, 18]. The base-optimizer explores multiple inner steps to find weight-updating directions. Then the meta-optimizer updates the meta-parameters following the explored trends. After few epochs of bilevel training, we learn a mapping function θ^{prior} from image to depth. It becomes better initialization for the subsequent supervised learning (Fig. 1). Note that meta-learning and the following supervised learning operate on the same training set without using extra data.

We show that meta-learning induces a *prior* with **higher robustness to unseen datasets**. To validate gain brought by meta-learning, we adopt multiple popular indoor datasets [19, 23, 27, 28] and devise protocols for **zero-shot cross-dataset evaluation**. This greatly differs from most previous works focusing only on intra-dataset evaluation, training and testing on a single dataset.

We first operate on limited data variety where meta-learning has at most 29.4% improvements, benefited by meta-learning's few-shot advantages. Then we go beyond limited scenes and train on datasets of a wide variety to validate our performance gain on larger-scale datasets for practical applications. We qualitatively and quantitatively show consistently superior performance by meta-learning on various network structures.

Contributions:

- The first method to apply meta-learning on pure single-image depth estimation to gain higher robustness without using additional training data, side information, or pre-trained networks.
- A novel fine-grained task concept in meta-learning to overcome the challenging single-image setting without obvious task boundaries. This becomes an empirical study for a complicated and practical target in meta-learning.
- Extensive experiments of zero-shot cross-dataset evaluation of indoor scenes to faithfully evaluate a model's robustness and generalizability, and results show consistently better performance using the meta-initialization strategy.

2. Methods

2.1. Limited-Data Monocular Depth Estimation

A model needs to distinguish depth-relevant and depth-irrelevant low-level cues for accurate estimation. The former shows color or radiance changes at object boundaries, and for the latter, geometry is invariant to color changes, such as decoration or object textures.

Limited-data impose challenges for robust estimation. It heavily relies on sufficient **scene variety** in training data, which demonstrate mappings from images to depth and enable learning from global context to suppress high-frequency depth-irrelevant cues. To gain robustness without using extra data, we exploit meta-learning's few-shot advantage and attain higher generalizability. Then we propose fine-grained task to adapt meta-learning to single-image depth estimation.

2.2. Single RGB-D Pair as Fine-Grained Task

Definition. Single-image depth prediction learns a function $f_\theta : \mathcal{I} \rightarrow \mathcal{D}$, parameterized by θ, to map from imagery to depth. A training set ($\mathbb{I}_{train}$, $\mathbb{D}_{train}$), containing image $I \in \mathbb{I}_{train}$ and associated depth map $D \in \mathbb{D}_{train}$, is used to train a model. In a minibatch with size K, each pair (I_i, D_i), $\forall i \in [1, K]$ is treated as a **fine-grained task**. Fine-grained tasks are mutual-exclusive: no two scenes sampled from the meta-distribution, i.e., the whole RGB-D dataset, share the same scene appearance and depth relation. Proof: assume we have two different scene images I_1 and I_2, and each contains sets of regions $\mathbb{R}_1$ and $\mathbb{R}_2$. The null set $\phi \notin \mathbb{R}^-$, where $\mathbb{R}^- = (\mathbb{R}_1 - \mathbb{R}_2) \cup (\mathbb{R}_2 - \mathbb{R}_1)$ that contains regions appear only in either I_1 or I_2, since I_1 and I_2 are different frames and inevitably capture different regions. Thus, any two scenes have different appearance and depth relations.

Difference with task in meta-learning. Fine-grained tasks are different from tasks in most-used meta-learning or few-shot learning usages [3], where a task contains data distribution and batches are sampled from it. Fine-grained tasks do not contain data distribution but are sampled from meta-distribution, the whole RGB-D dataset. For example, a navigating agent captures image and depth pairs. The RGB-D pairs are sampled from the meta-distribution.

Design. Each fine-grained task is used to learn on the specific RGB-D pair. The design is motivated by the fact that appearance and depth variation can be high. A view looking at small desk objects and a view of large room spaces are highly dissimilar in contents and ranges. Mappings from their scene appearance to range values are different. Still, they can be captured in the same environment or even in neighboring frames. This contrasts with image classification where class samples share a common label. The observation explains why we treat each RGB-D pair as a fine-grained task instead of each environment.

2.3. Meta-Initialization on Depth from Single Image

We describe our approach based on gradient-based meta-learning to learn a good initialization (Fig. 1).

Prior learning stage. In the first prior learning stage, we adopt a meta-optimizer and a base-optimizer. In each meta-iteration, K fine-grained tasks as a minibatch are sampled from the whole training set: $(I_i, D_i) \sim (\mathbb{I}_{train}, \mathbb{D}_{train})$, $\forall i \in [1, K]$. Then we take L steps to explore gradient directions that minimize the regression loss and get $(\theta^1_{expl}, \theta^2_{expl}, ..., \theta^L_{expl})$:

$$\theta^i_{expl} \leftarrow \theta^{i-1}_{expl} - \alpha \frac{1}{K} \nabla_\theta \sum_{k \in [1,K]} \mathcal{L}_{reg}(I_k, D_k; \theta^{i-1}_{expl}), \forall i \in [1, L]. \tag{1}$$

After the L-step exploration, we update the meta-parameters using Reptile style [18], i.e., following the explored weight updating direction in the inner steps.

$$\theta^j_{meta} \leftarrow \theta^{j-1}_{meta} - \beta(\theta^{j-1}_{meta} - \theta^L_{expl}), \tag{2}$$

where α and β are respective learning rates. i and j denote inner and meta-iterations.

Compared with MAML [3], we find Reptile is more suitable for training for fine-grained tasks. First, as mentioned in Reptile's paper [18], it is designed without support and query set split, and thus it inherently does not require multiple data samples in a task, which matches our fine-grained task definition. Next, first-order MAML computes gradients on the query set at the last inner step θ^L_{expl} to update meta-parameters. However, only one sample exists in each fine-grained task, and each fine-grained task is mutual-exclusive and can differ greatly, depending on $\mathbb{R}^-$. Thus, if taking exploration on a support split and computing gradients on the query split, but the support and query samples do not share common components, the gradients are nearly random and prevent from converging. By contrast, Reptile does not entail support and query split or require common components between samples, so it stabilizes training towards convergence and becomes the choice.

Supervised learning stage. Prior knowledge θ^{prior} is learned after the first stage. We treat it as the initialization for the subsequent supervised learning with conventional stochastic gradient descent to minimize the regression loss.

$$\theta^* \leftarrow \min_\theta \mathcal{L}_{reg}(\mathbb{I}_{train}, \mathbb{D}_{train} | \theta^{prior}). \tag{3}$$

Last, test set ($\mathbb{I}_{test}, \mathbb{D}_{test}$) is used to evaluate performance of θ^*. We show the pseudo-code and organized algorithm in Supplementary (Supp) Sec.G. The implementation only needs a few-line codes as plugins to depth estimation frameworks, which benefits higher model generalizability as shown in later experiments.

Difference with other learning strategies. Compared with widely-used pretraining that requires multiple data sources to gain generalizability [21, 22, 29], both the prior learning and supervised learning stages operate on the same dataset

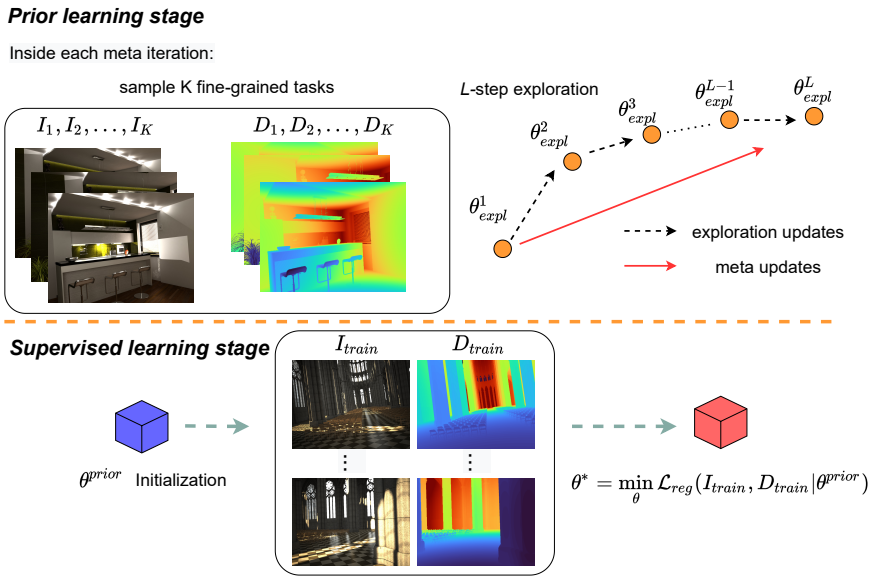

Figure 1. **Meta-Initialization for learning image-to-depth mappings.** The prior learning stage adopts a base-optimizer and a meta-optimizer. Inside each meta-iteration, K fine-grained tasks are sampled and used to minimize regression loss. L steps are taken by the base-optimizer to search for weight update directions for these K tasks. Then, the meta-optimizer follows the explored inner trends to update meta-parameters in the Reptile style [18]. Image-to-depth prior θ^{prior} is output at the end of the stage. θ^{prior} is then used as the initialization for the subsequent supervised learning for the final model θ^*.

without access to extra data or off-the-shelf models. Thus, they are free from those burdens.

Compared with simple gradient accumulation [24], where gradients are accumulated for several batches and then used to update parameters only once, the bilevel optimization keeps updating the inner-parameters every step in L to find the local niche for the current batch. Besides, gradient accumulation has the effect of large batch size, which might cause overfitting and degrade model generalizability.

2.4. Strategy and Explanation

Meta-Initialization. We next analyze meta-learning behavior with fine-grained task. Inside each meta-iteration, the base-learner explores the neighborhood with L-step using K fine-grained tasks. Compared with simple single-step update, the meta-update can be seen as first taking L-step amortized gradient descent with a lower learning rate to delicately explore local loss manifolds, then updating meta-parameters by trends shown in the inner steps with a step size β towards θ^L. θ^{prior} after the prior learning may underfit the training set since the algorithm suggests not wholly following optimal gradients for each batch but with a β for control. However, it avoids overfitting to seen RGB-D pairs and forces the inner exploration to reach higher-level image-to-depth understanding. θ^{prior} then becomes good initialization for downstream RGB-D learning.

Progressive learning perspective. Meta-initialization can be seen as progressive learning on a training set. At the first stage, meta-learning benefits learning coarse but smooth depth from global context. In Supp Fig. S3 we apply the first-stage meta-learning compared to direct supervised learning on a dataset of limited scene variety. Meta-learning estimates smooth depth shapes and is free from irregularity that direct supervised learning encounters. The irregularity indicates the dataset did not provide sufficient scene variety that demonstrates how images map to depth in various environments to learn smooth depth from global context. Consequently, only local high-frequency cues show. To illustrate, if only sparse and irrelevant scene images are presented,

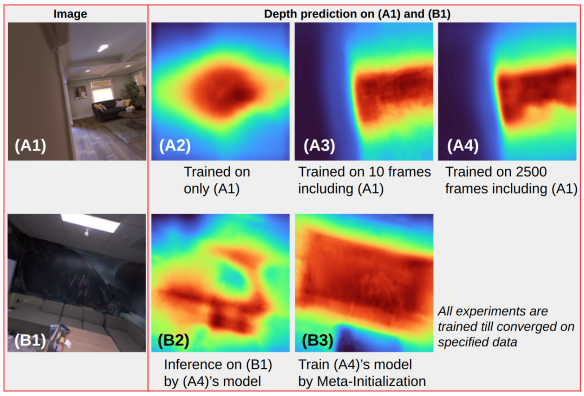

Figure 2. **Analysis on scene variety and model generalizability.** (A) shows limited training scenes constrain learning image-to-depth mappings, with an extreme case (A2) for only one training image. (B) shows though a model (A4) fits well on training scenes, it still cannot generalize to unseen seen, especially wall paintings with many depth-irrelevant cues. Meta-initialization attains better model generalizability.

Table 1. **Generalizability with different scene variety.** We compare single-stage meta-learning (Meta) and direct supervised learning (DSL) using ConvNeXt-Base. $\rightarrow$ specifies train and test datasets. Replica and HM3D respectively hold lower and higher scene variety for training. Meta-Learning has much larger improvements especially trained on low scene-variety Replica.

	Replica $\rightarrow$ VA			HM3D $\rightarrow$ VA		
Method	MAE	AbsRel	RMSE	MAE	AbsRel	RMSE
DSL	0.718	0.538	1.078	0.544	0.456	0.715
Meta	**0.548**	**0.430**	**0.761**	**0.427**	**0.369**	**0.603**
	-23.6%	*-20.1%*	*-29.4%*	*-21.5%*	*-19.1%*	*-15.7%*

finding a function that satisfactorily fits those scenes with smooth depth from global context is hard. See Fig. 2. The irregularity occurs especially at cluttered objects or surface textured areas, since those *depth-irrelevant local cues are barely suppressed*. In summary, the progressive fashion first learns coarse but smooth depth by θ^{prior}. Then, the network learns finer depth at the second supervised stage.

3. Experiments

Aims. We validate our meta-initialization with the questions. **Q1** Can meta-learning improve learning image-to-depth mapping on limited scene-variety datasets? (Sec. 3.1) **Q2** How does meta-initialization help zero-shot cross-dataset generalization? (Sec. 3.2) We use ResNet [5] and ConvNeXt [17] as backbones. See Supp Sec.F for training settings.

Datasets: We adopt high scene-variety HM3D [19] and Hypersim [23] as training sets and use Replica [27], VA [28], NYUv2 [26]as test sets. Details, exemplar data, and evaluation metrics are given in Supp Sec.B and F.

3.1. Meta-Learning on Limited Scene Variety

We first show how a single-stage meta-learning (only prior learning) performs. We train N=15 epochs of meta-learning and compare with 15 epochs of direct supervised learning where both training pieces converge already. Replica Dataset of limited scene variety is used to verify gain on limited sources. We numerically show generalizability to unseen datasets. HM3D (high scene variety) and Replica (low scene variety) are used as training sets and VA is used for testing. Table 1 shows that models trained by single-stage meta-learning substantially outperform direct supervised learning with 15.7%-29.4% improvements. The advantage is more evident when trained on lower scene-variety Replica, which matches meta-learning's advantages on few-shot learning.

3.2. Meta-Initialization on Higher Scene Variety

We next train full meta-initialization algorithm. In the section, **go beyond limited sources** and train on higher scene-variety datasets. Intuitively, higher scene variety helps supervised learning attain better depth prediction and might diminish meta-learning's advantages of few-shot and low-source learning. However, such studies are practical for validating meta-learning in real-world applications. High scene-variety and larger-size synthetic datasets, Hypersim and HM3D, are used as training sets. VA, Replica, and NYUv2 serve as testing, and their evaluations are capped at 10m. We median-scale prediction to groundtruth in the protocol to compensate for different camera intrinsic.

Table 2. **Zero-Shot cross-dataset evaluation using meta-initialization.** Comparison is drawn between without meta-initialization (no marks, ImageNet-initialization) and with our meta-initialization (Meta) using different sizes ConvNeXt. Results of "+Meta" are consistently better.

Hypersim $\rightarrow$ VA	MAE	AbsRel	RMSE	δ_1	δ_2	δ_3
ConvNeXt-small	0.291	0.215	0.404	68.5	90.8	96.7
ConvNeXt-small + Meta	**0.280**	**0.207**	**0.398**	**70.4**	**91.3**	**97.0**
ConvNeXt-base	0.275	0.201	0.393	71.3	91.8	97.3
ConvNeXt-base + Meta	**0.259**	**0.194**	**0.365**	**72.8**	**92.8**	**97.8**
ConvNeXt-large	0.263	0.198	0.369	73.0	92.0	97.1
ConvNeXt-large + Meta	**0.248**	**0.183**	**0.355**	**74.6**	**93.5**	**97.8**
Hypersim $\rightarrow$ NYUv2	MAE	AbsRel	RMSE	δ_1	δ_2	δ_3
ConvNeXt-small	0.434	0.165	0.598	75.7	94.3	98.5
ConvNeXt-small + Meta	**0.415**	**0.155**	**0.575**	**77.8**	**95.1**	**98.8**
ConvNeXt-base	0.396	0.150	0.549	79.6	95.6	98.9
ConvNeXt-base + Meta	**0.386**	**0.141**	**0.524**	**80.3**	**96.0**	**99.0**
ConvNeXt-large	0.389	0.149	0.542	79.8	95.6	98.8
ConvNeXt-large + Meta	**0.375**	**0.140**	**0.517**	**81.2**	**96.2**	**99.1**
Hypersim $\rightarrow$ Replica	MAE	AbsRel	RMSE	δ_1	δ_2	δ_3
ConvNeXt-small	0.307	0.189	0.417	72.4	92.1	**97.5**
ConvNeXt-small + Meta	**0.294**	**0.178**	**0.404**	**74.5**	**92.7**	**97.5**
ConvNeXt-base	0.312	0.185	0.429	74.1	92.6	97.4
ConvNeXt-base + Meta	**0.288**	**0.173**	**0.399**	**75.6**	**93.3**	**97.9**
ConvNeXt-large	0.285	0.172	0.394	75.8	93.2	97.7
ConvNeXt-large + Meta	**0.273**	**0.165**	**0.380**	**77.0**	**94.0**	**98.1**

In Table 2, compared with ImageNet-initialization, meta-initialization **consistently** improves in nearly all the metrics, especially δ_1 (averagely +1.97 points). The gain comes from that meta-prior attains a better image-to-depth mapping. Conditioned on the initialization, the learning better calibrates to open-world image-to-depth relation hence generalizes better to unseen scenes. We further apply meta-initialization to dedicated architecture for depth and show consistently better results. See Supp Sec. H for more qualitative and quantitative results and studies on depth-supervised NeRF and Supp. Sec. C for detailed related work.

4. Conclusion and Limitation

From depth's perspective, this work studies a learning scheme to gain robustness without extra data or constraints. We further propose a zero-shot cross-dataset protocol to attend to in-the-wild robustness that most prior works overlook. From meta-learnings's view, we propose fine-grained task to overcome the lacks of affinity in sparse and irrelevant sampled images. Further we study a complex single-image real-valued regression problem rather than widely-studied classification.

The work is at the intersection of the two research fields and we hope it drives the dual-stream research development.

References

[1] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Localbins: Improving depth estimation by learning local distributions. In *ECCV*, 2022. 1

[2] Jia-Wang Bian, Huangying Zhan, Naiyan Wang, Tat-Jun Chin, Chunhua Shen, and Ian Reid. Auto-rectify network for unsupervised indoor depth estimation. *TPAMI*, 2021. 1

[3] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, 2017. 1, 2

[4] Chelsea Finn, Aravind Rajeswaran, Sham Kakade, and Sergey Levine. Online meta-learning. In *ICML*, 2019. 1

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 4

[6] Xu He, Jakub Sygnowski, Alexandre Galashov, Andrei A Rusu, Yee Whye Teh, and Razvan Pascanu. Task agnostic continual learning via meta learning. *arXiv preprint arXiv:1906.05201*, 2019. 1

[7] Sepp Hochreiter, A Steven Younger, and Peter R Conwell. Learning to learn using gradient descent. In *ICANN*, 2001. 1

[8] Aleksander Holynski and Johannes Kopf. Fast depth densification for occlusion-aware augmented reality. In *SIGGRAPH Asia*, 2018. 1

[9] Timothy M Hospedales, Antreas Antoniou, Paul Micaelli, and Amos J Storkey. Meta-learning in neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 1

[10] Pan Ji, Runze Li, Bir Bhanu, and Yi Xu. Monoindoor: Towards good practice of self-supervised monocular depth estimation for indoor environments. In *ICCV*, 2021. 1

[11] Hualie Jiang, Laiyan Ding, Junjie Hu, and Rui Huang. Plnet: Plane and line priors for unsupervised indoor depth estimation. In *3DV*, 2021. 1

[12] Jinyoung Jun, Jae-Han Lee, Chul Lee, and Chang-Su Kim. Depth map decomposition for monocular depth estimation. 2022. 1

[13] Doyeon Kim, Woonghyun Ga, Pyungwhan Ahn, Donggyu Joo, Sehwan Chun, and Junmo Kim. Global-local path networks for monocular depth estimation with vertical cutdepth. *arXiv preprint arXiv:2201.07436*, 2022. 1

[14] Boying Li, Yuan Huang, Zeyu Liu, Danping Zou, and Wenxian Yu. Structdepth: Leveraging the structural regularities for self-supervised indoor depth estimation. In *ICCV*, 2021. 1

[15] Zhenyu Li, Zehui Chen, Xianming Liu, and Junjun Jiang. Depthformer: Depthformer: Exploiting long-range correlation and local information for accurate monocular depth estimation. *arXiv preprint arXiv:2203.14211*, 2022. 1

[16] Zhenyu Li, Xuyang Wang, Xianming Liu, and Junjun Jiang. Binsformer: Revisiting adaptive bins for monocular depth estimation. 2022. 1

[17] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. *CVPR*, 2022. 4

[18] Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*, 2018. 1, 2, 3

[19] Santhosh K Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alex Clegg, John Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X Chang, et al. Habitat-matterport 3D dataset (HM3D): 1000 large-scale 3D environments for embodied AI. *NeurIPS Datasets and Benchmarks Track*, 2021. 1, 4

[20] Michael Ramamonjisoa and Vincent Lepetit. Sharpnet: Fast and accurate recovery of occluding contours in monocular depth estimation. *ICCVW*, 2019. 1

[21] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. *ICCV*, 2021. 2

[22] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *TPAMI*, 2020. 2

[23] Mike Roberts and Nathan Paczan. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. *ICCV*, 2021. 1, 4

[24] Sebastian Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016. 3

[25] Jürgen Schmidhuber. Evolutionary principles in self-referential learning, or on learning how to learn: the meta-meta-... hook. *Technische Universität München, PhD thesis*, 1987. 1

[26] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 4

[27] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, et al. The replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 1, 4

[28] Cho-Ying Wu, Jialiang Wang, Michael Hall, Ulrich Neumann, and Shuochen Su. Toward practical monocular indoor depth estimation. In *CVPR*, 2022. 1, 4

[29] Wei Yin, Jianming Zhang, Oliver Wang, Simon Niklaus, Long Mai, Simon Chen, and Chunhua Shen. Learning to recover 3d scene shape from a single image. In *CVPR*, 2021. 2

[30] Zehao Yu, Lei Jin, and Shenghua Gao. P^2net: Patch-match and plane-regularization for unsupervised indoor depth estimation. In *ECCV*, 2020. 1

[31] Weihao Yuan, Xiaodong Gu, Zuozhuo Dai, Siyu Zhu, and Ping Tan. New crfs: Neural window fully-connected crfs for monocular depth estimation. *CVPR*, 2022. 1

[32] Junsheng Zhou, Yuwang Wang, Kaihuai Qin, and Wenjun Zeng. Moving indoor: Unsupervised video depth learning in challenging environments. In *CVPR*, 2019. 1